|  |
| --- |
| Informe lab 12  Big Data |
| Integrantes:   * Cristian Álvarez * Jorge González |



# Introducción

En este informe se documentará todo lo realizado y solicitado en el “Lab12”, el cual se divide en dos partes, la primera de ellas es la documentación del proceso en que se hace la construcción de una maquina virtual con Docker incorporado para así instalar Hadoop de Cloudera, mongoDB y Spark; la segunda parte es explicar 4 puntos de vista de un artículo específico que se nos indica en el laboratorio.

# Desarrollo

#### Instrucciones

**Parte1**

* Realice el trabajo en grupos de 2 estudiantes.
* Construya una máquina virtual (MV) en VirtualBox con Docker e instale Hadoop de Cloudera, mongoDB y Spark.
* La MV tendrá un único usuario aparte del root, el nombre de este usuario será el apellido de uno de los estudiantes del grupo.
* En un informe de máximo 5 páginas, muestre el proceso de instalación y verifique que las aplicaciones están ejecutándose en el sistema.
* La máquina virtual será revisada en Laboratorio o será evaluada con nota mínima. Si es necesario debe traerla en Pendrive.

**Parte2**

* Lea el siguiente artículo y explique 4 puntos de vistas desde él en máximo 2 hojas del informe.

<https://www.kdnuggets.com/2017/04/hadoop-not-failing-future-data.html>

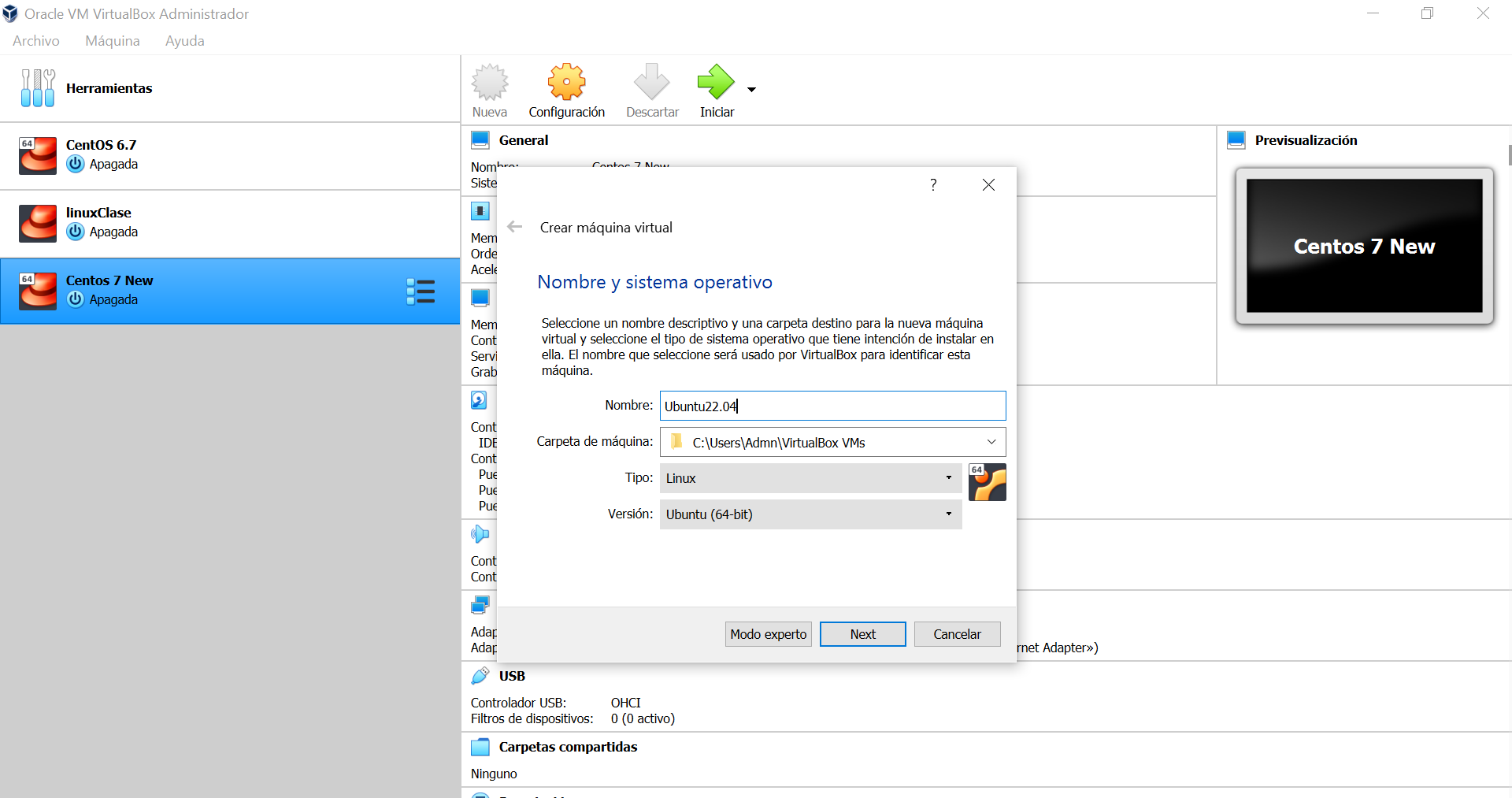
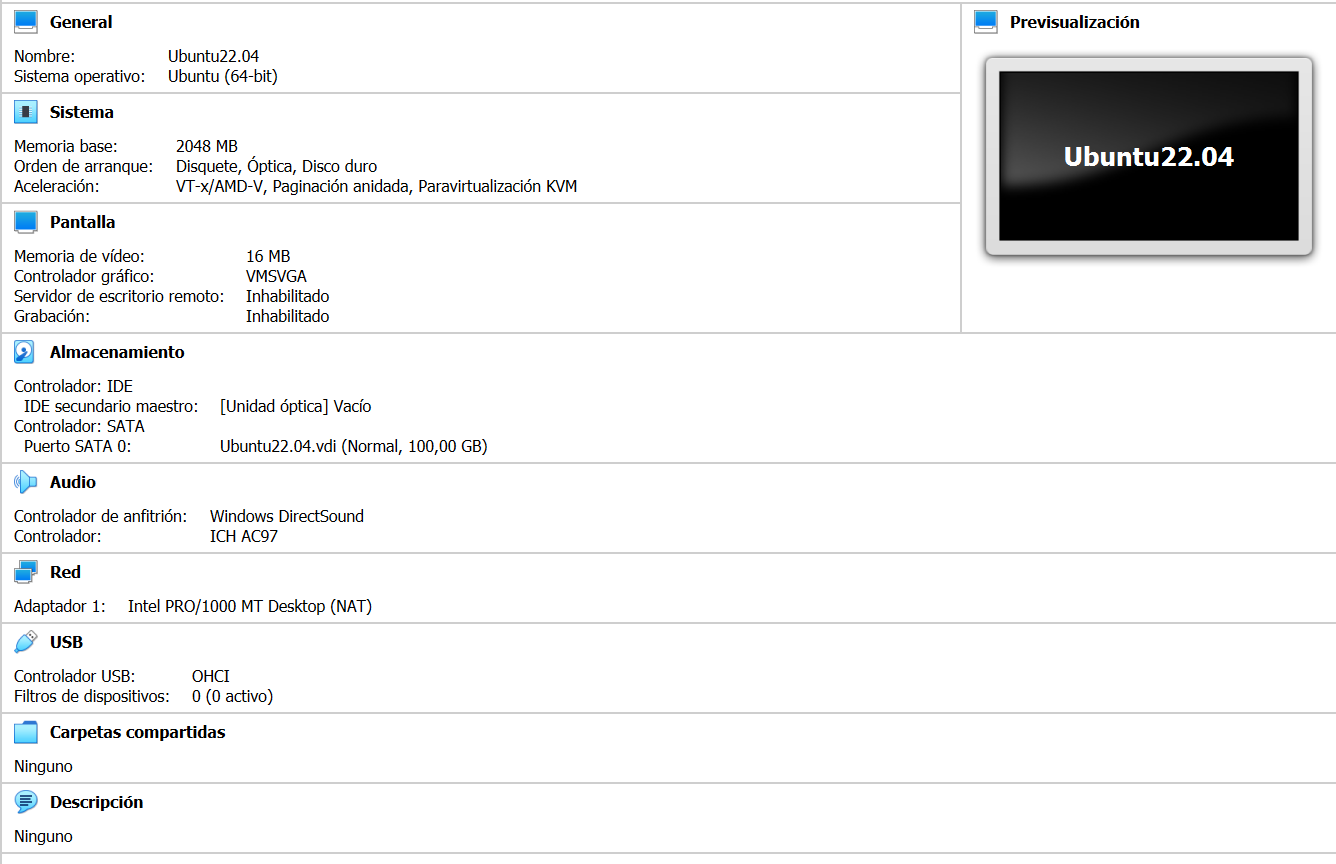
Lab12 Parte1

Dado que existen gran cantidad de pasos en cada procedimiento, las imágenes que representan cada paso están en un tamaño reducido, pero aun así haciendo Zoom en el informe podrá ver con buena resolución la imagen, por ello ***recomiendo hacer un Zoom del 250% aproximadamente para ver las imágenes una a una.***

#### Construcción de MV en VirtualBox

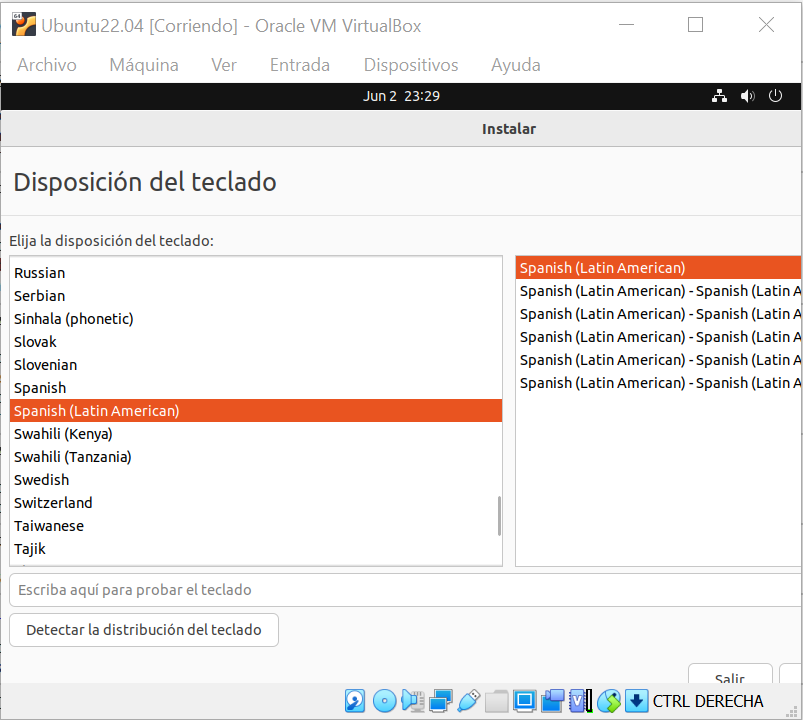
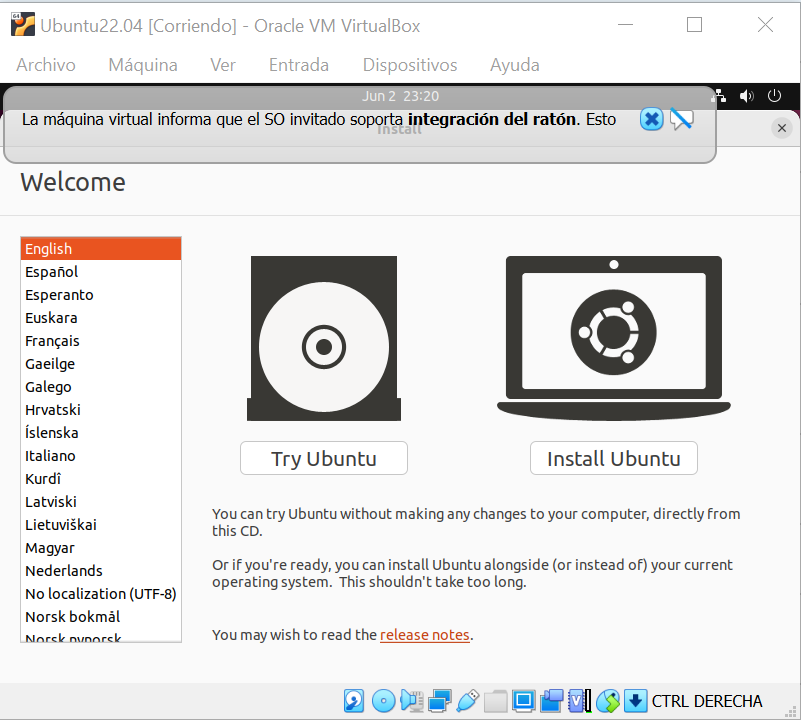
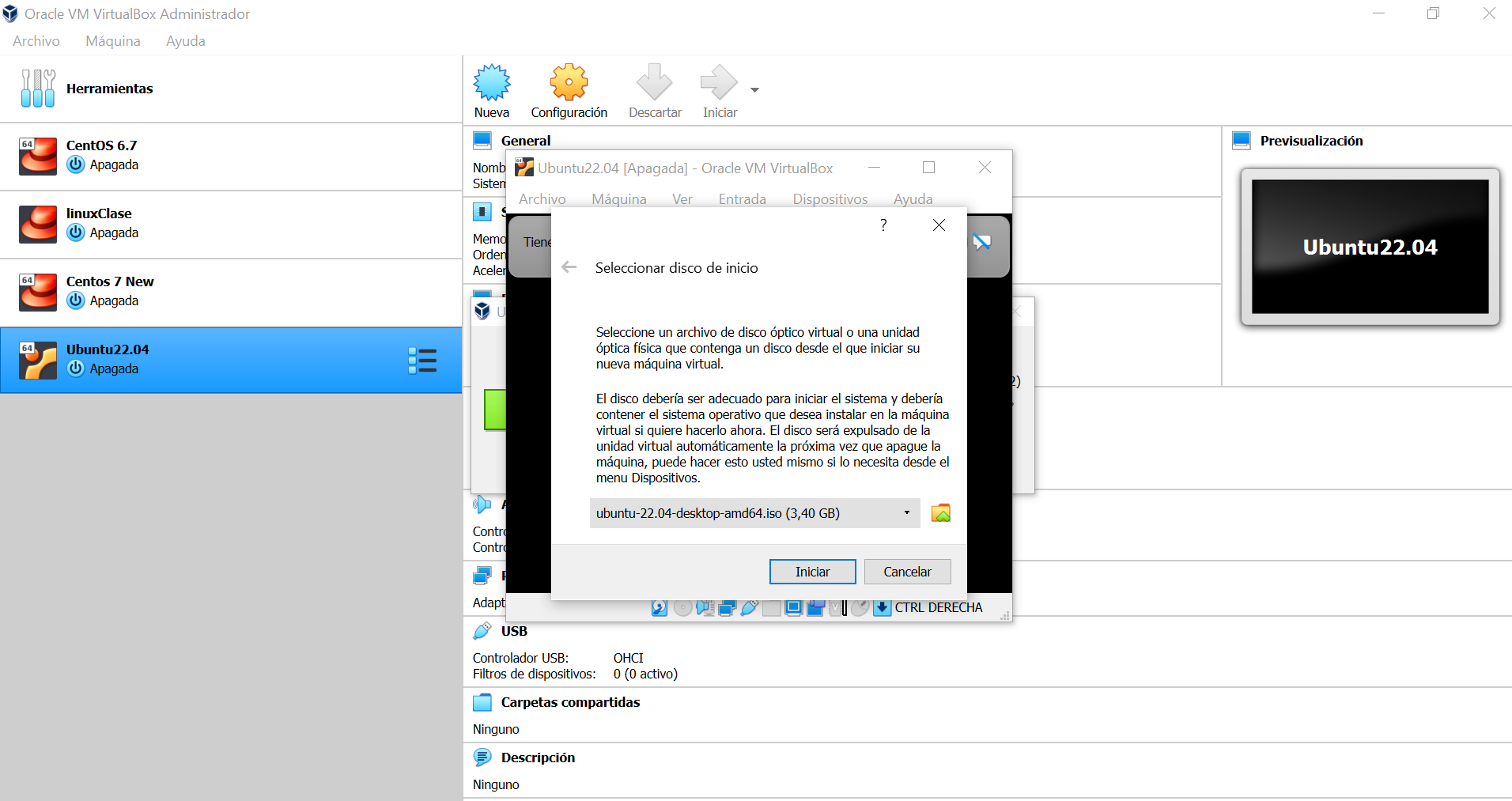
Para la construcción de la MV utilizamos el software de “Oracle VirtualBox”, donde utilizaremos la Imagen ISO de Ubuntu 22.04 y realizaremos su respectiva instalación.

##### Máquina Virtual y Características

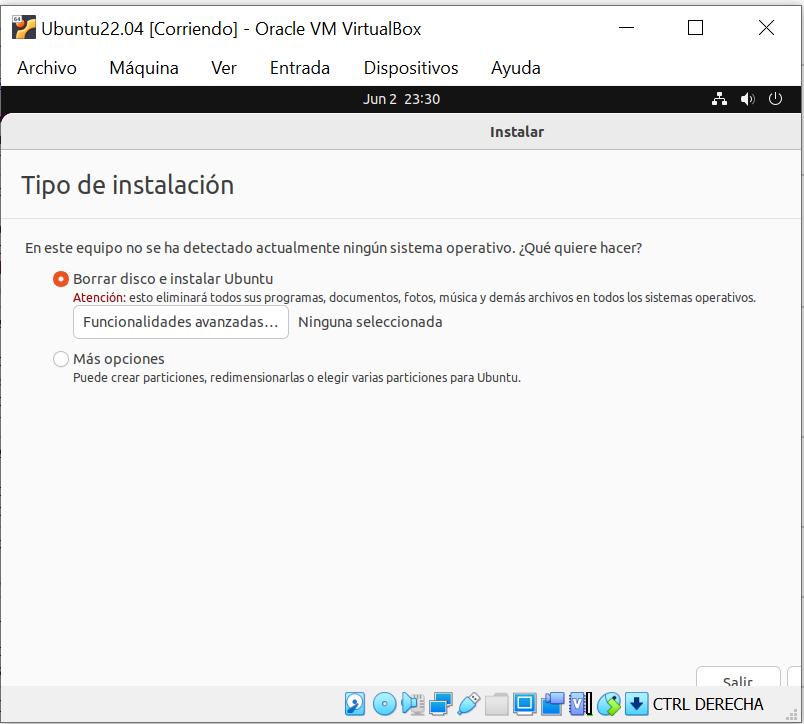
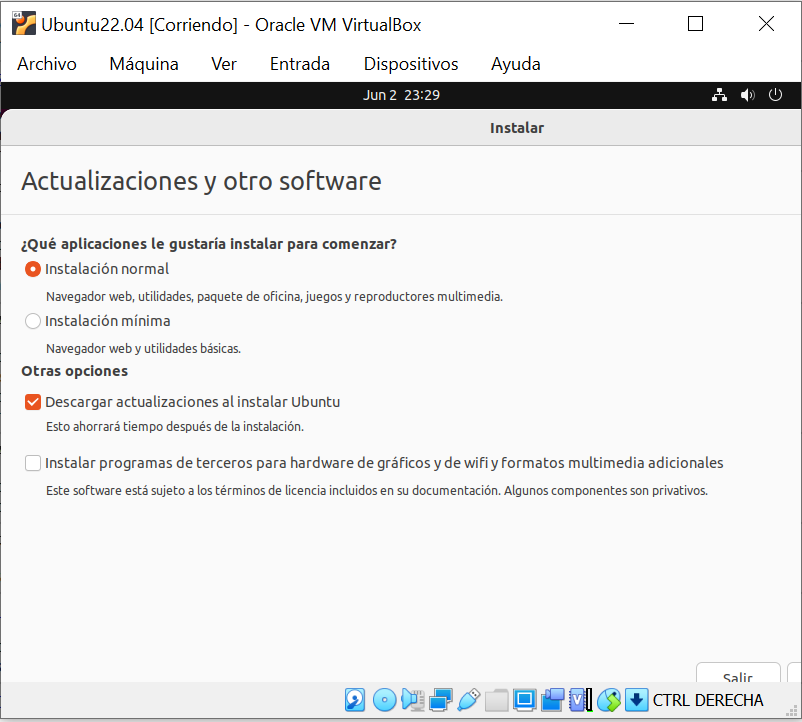
 

##### Carga e Instalación de Ubuntu 22.04

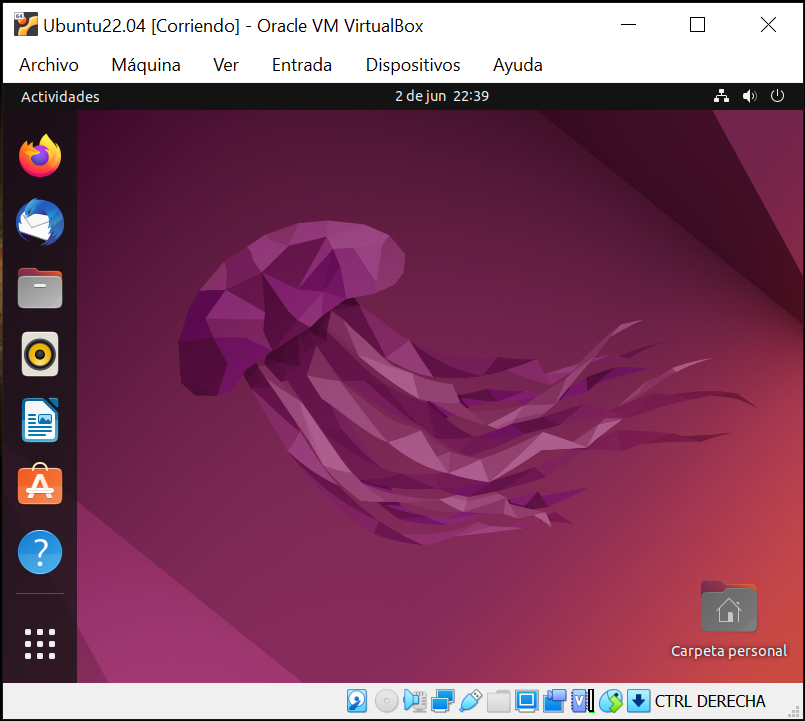
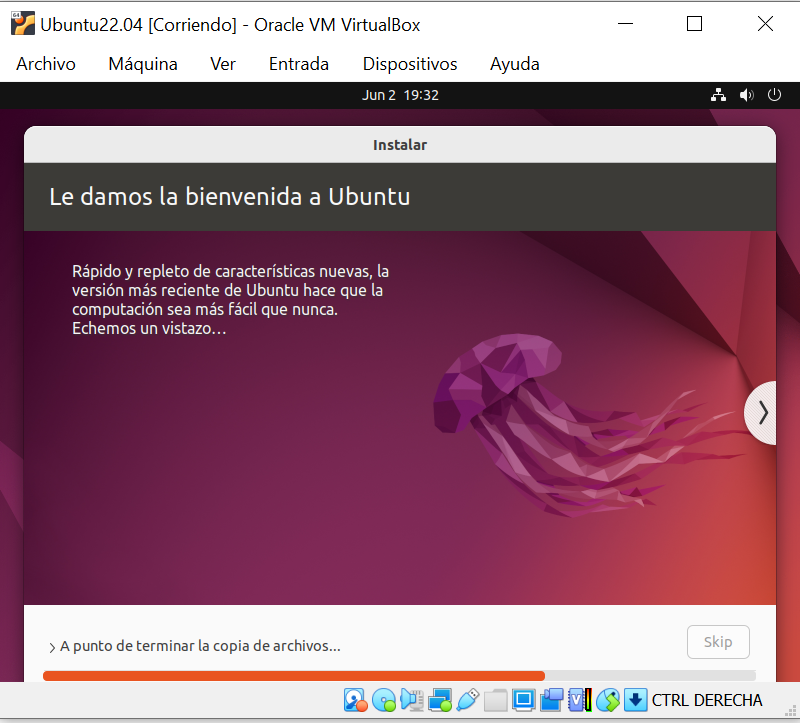
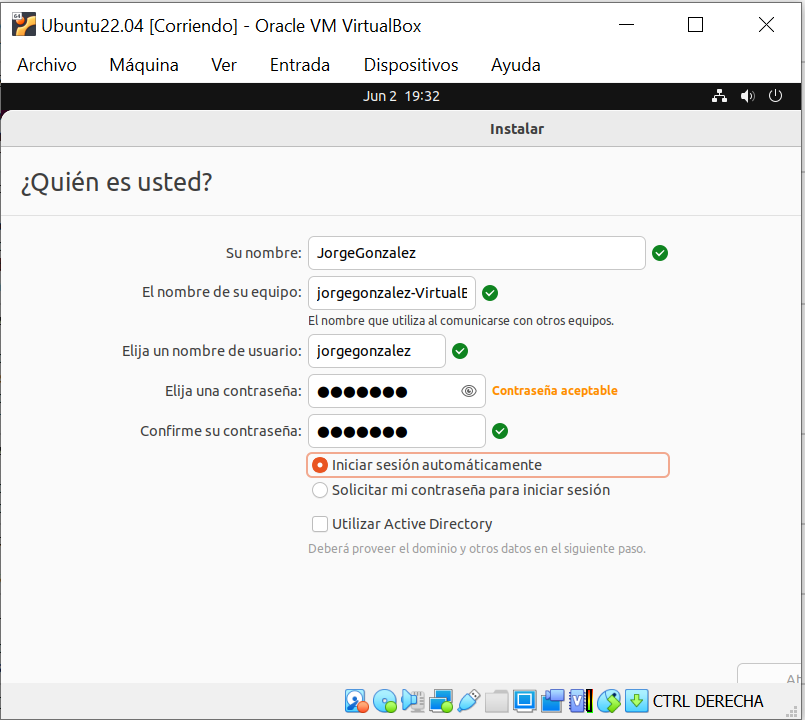
1. 2. 3.



4. 5. 6.



7. 8. 9.



##### Instalación de Docker en Ubuntu

1) -Ingresar el comando: sudo apt-get update

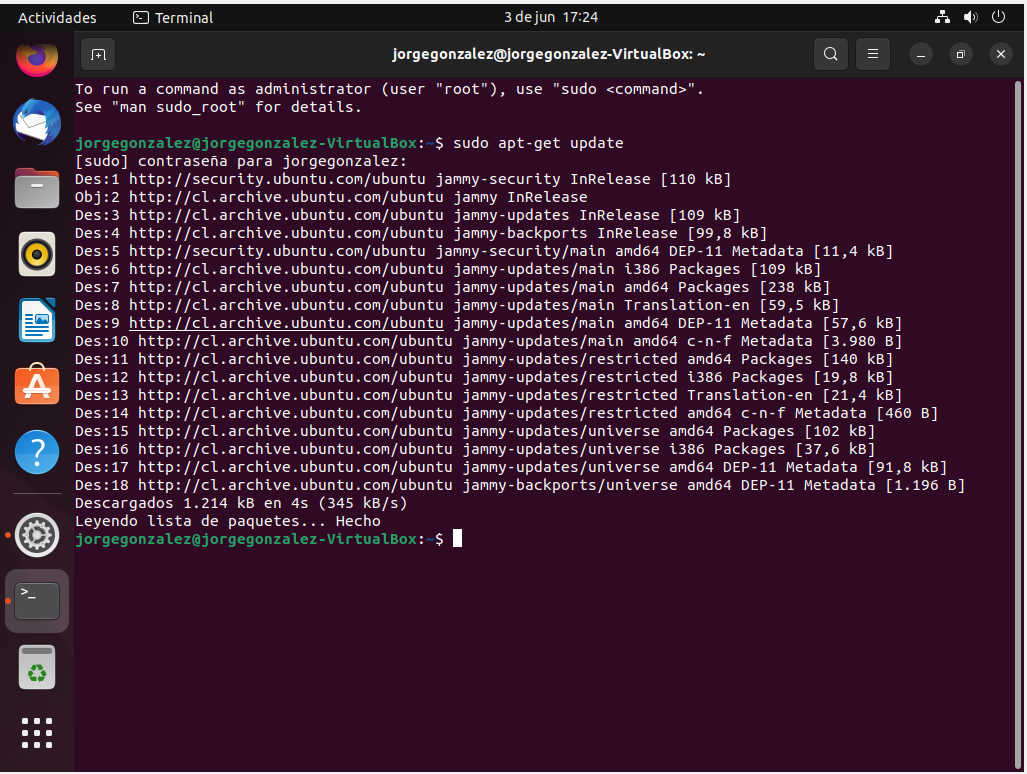
-Ingresar contraseña de usuario

2) -Luego la instalamos de unos complementos con el comando:

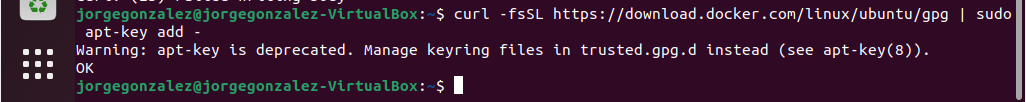
Sudo apt-get install apt-transport-https ca-certificates curl software-properties-common -y

3) -Ejecutamos el siguiente comando para validar los archivos con una key:

curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add –

*(Imagen punto 1) (Imagen punto 2)*



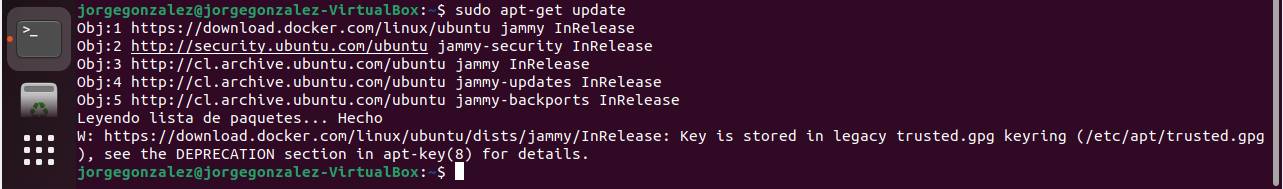
*(Imagen punto 3)*

4) -Agregamos el siguiente repositorio con el comando:

sudo add-apt-repository "deb [arch=amd64] https://download.docker.com/linux/ubuntu $(lsb\_release -cs) stable"

5) -Volvemos a realizar un Update:

sudo apt-get update

*(Imagen punto 4) (Imagen punto 5)*

6) -Realizamos la instalación directa de Docker

sudo apt-get install docker-ce -y

7) -Anclamos el inicio de Docker al iniciar el sistema

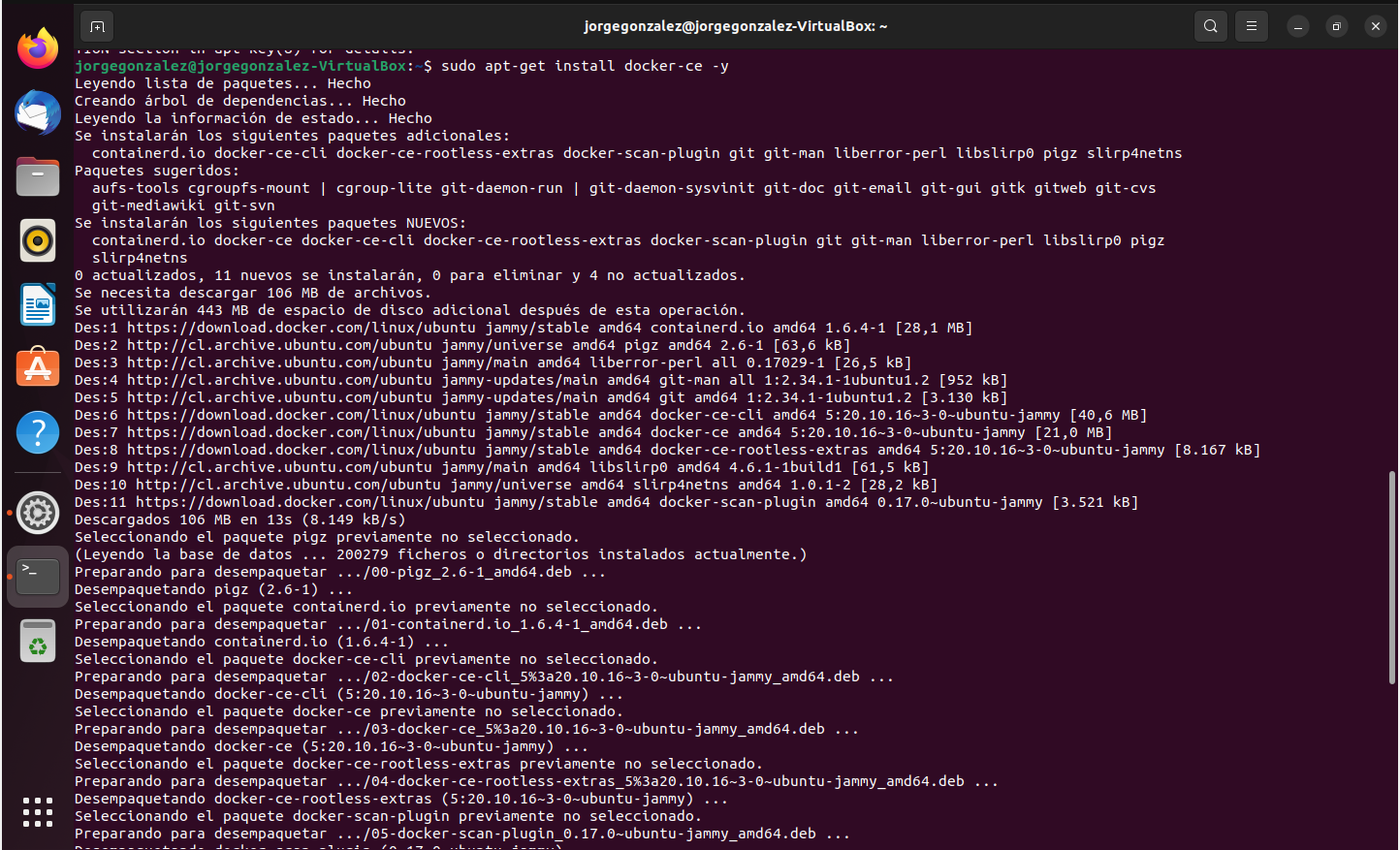
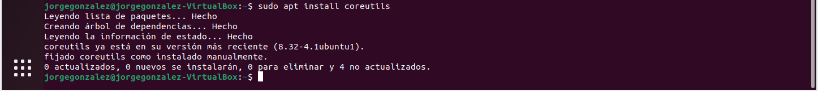
sudo systemctl enable docker

8) -Instalamos el paquete coreutils

sudo apt install coreutils

9) -Agregamos nuestro usuario a docker, proveniente de la salida del comando whoami

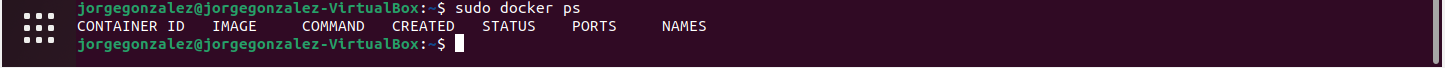
sudo usermod -aG docker (salida comando whoami)

 *(Imágenes puntos 7, 8, 9 respectivamente)*

*(Imagen punto 6)*

10) -Comprobación de instalación, ingresamos el siguiente comando:

sudo docker ps



*(Imagen punto 10)*

##### Instalación de Hadoop de Cloudera

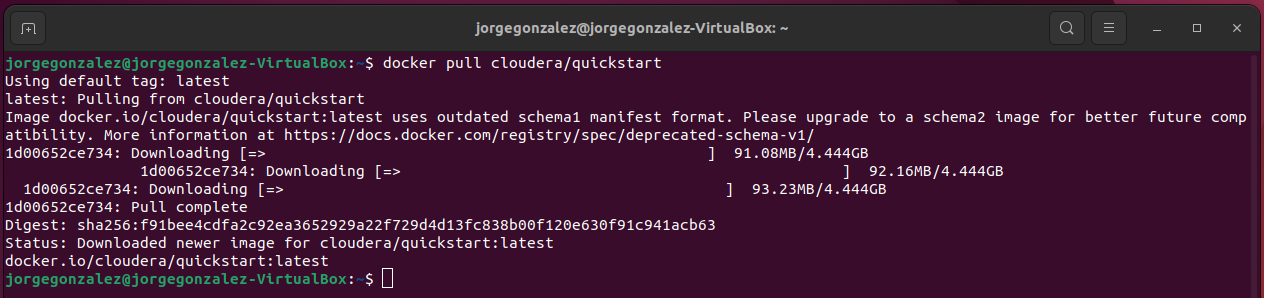
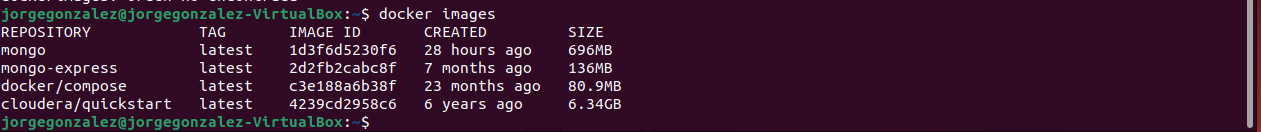
1. A través de docker hacemos la instalación de Cloudera/quickstart

docker pull cloudera/quickstart

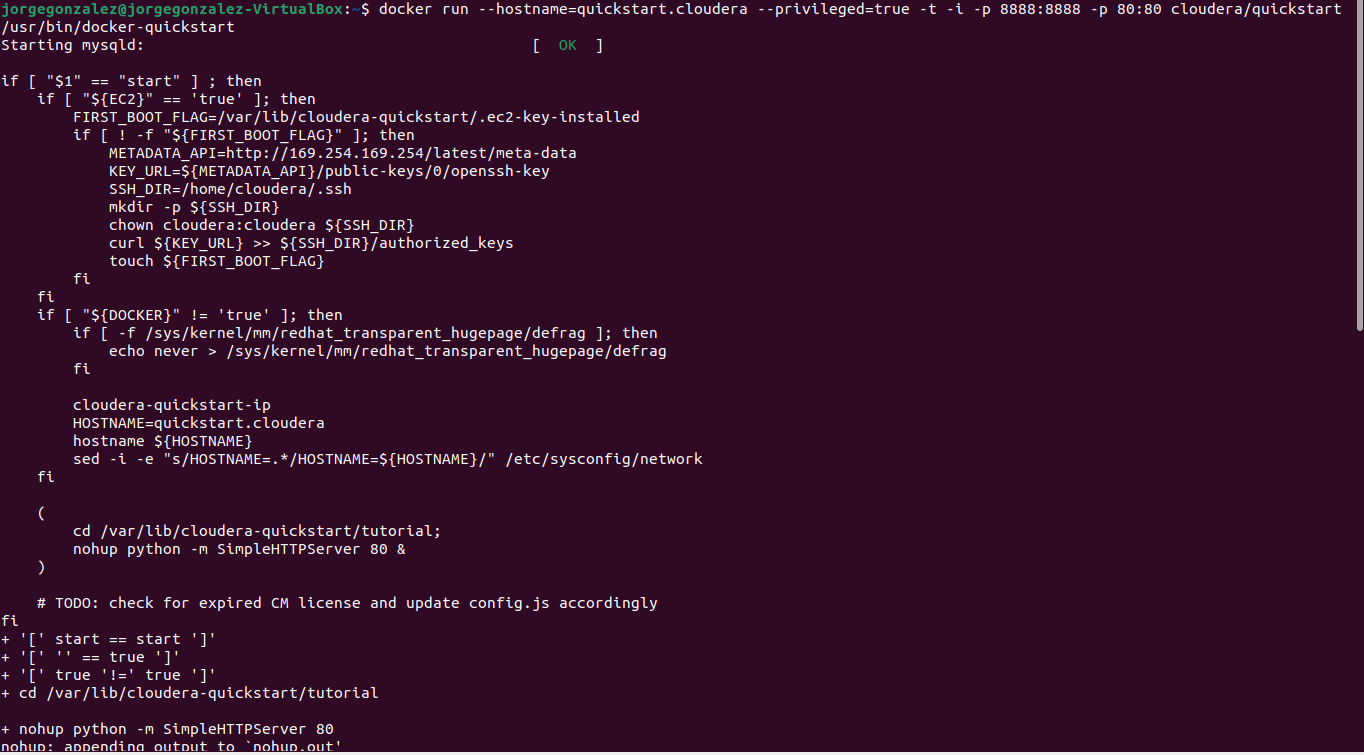
1. Comprobamos la existencia de la imagen para posteriormente hacerla correr

docker images

docker run --hostname=quickstart.cloudera --privileged=true -t -i -p 8888:8888 -p 80:80 cloudera/quickstart /usr/bin/docker-quickstart

*(Imagen punto 1) (Imagen punto 2.1)*



*(Imagen punto 2.2)*

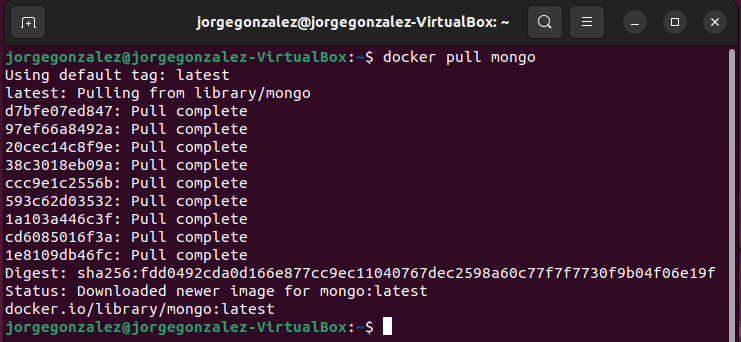
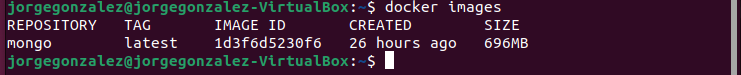
##### Instalación de MongoDB

1)Para la instalación de la imagen de Mongo en docker introducir el comando:

docker pull mongo

2)Comprobamos la instalación correcta de la imagen

docker image

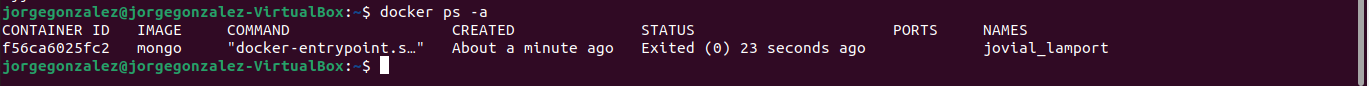
*(Imagen punto 1) (Imagen punto 2)*

3)Hacemos correr mongo para comprobar su funcionamiento

docker run mongo

4)Revisamos el “Historial de ejecución” de las imágenes y ver si se registró el run de mongo

docker ps -a

*(Imagen punto 3) (Imagen punto 4)*

##### Instalación de Spark

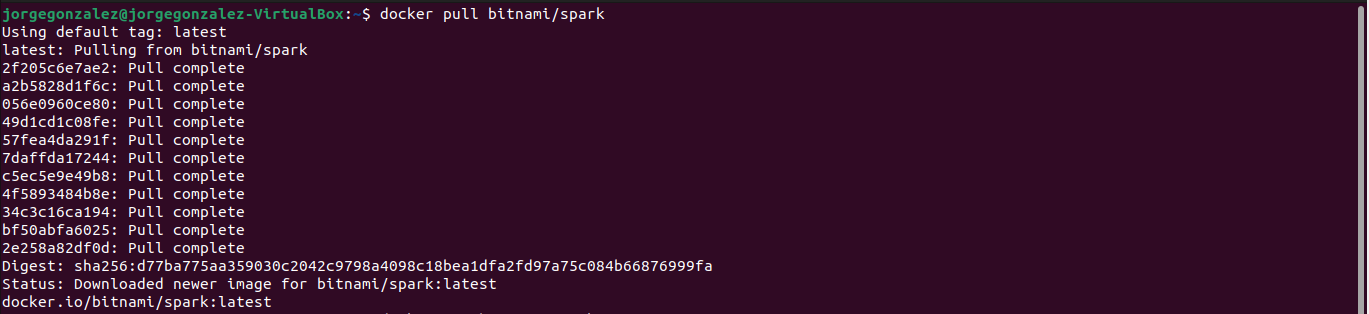
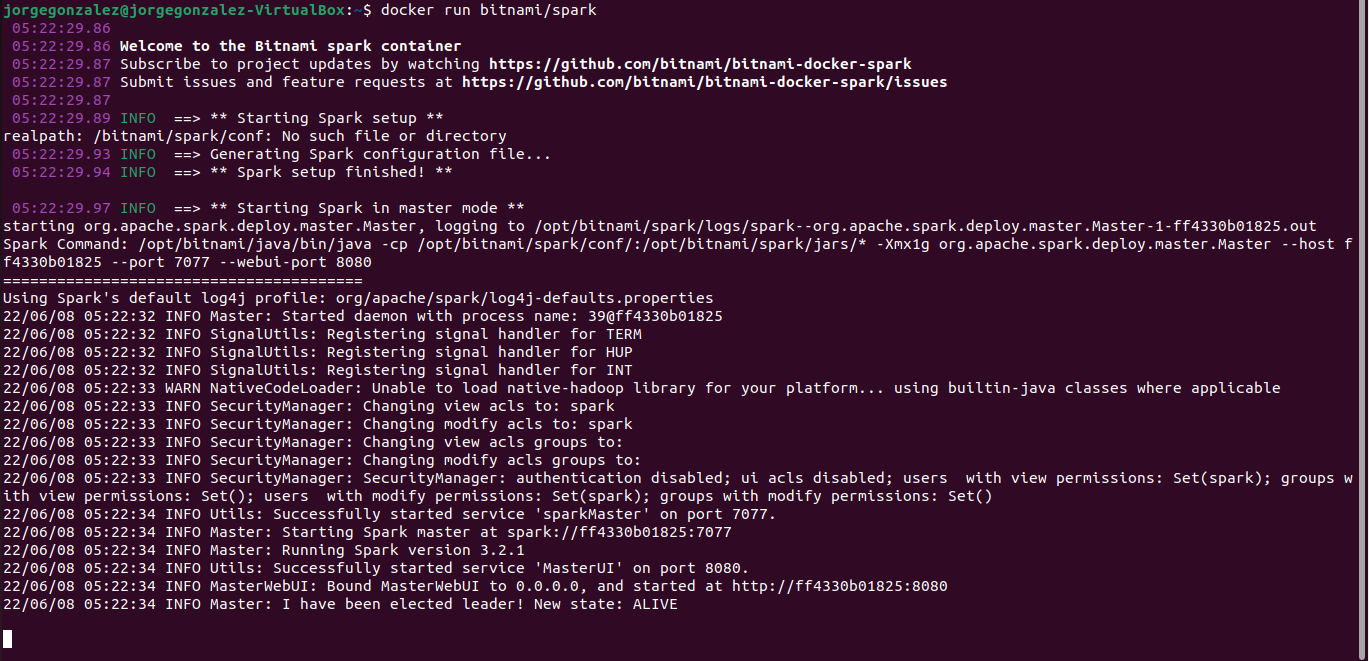
1. Se instala la imagen de Spark en docker con el comando:

docker pull bitnami/Spark

1. Se corre Spark con el comando respectivo y seguido vemos si existe la imagen.

docker run bitnami/Spark

docker images

*(Imagen punto 1) (Imagen punto 2)*



*(Imagen punto 3)*

Lab12 Parte2

#### Intrucciones

* Lea el siguiente artículo y explique 4 puntos de vistas desde él en máximo 2 hojas del informe.

<https://www.kdnuggets.com/2017/04/hadoop-not-failing-future-data.html>

***Punto1.*** En este articulo existe un punto de vista principal y del cual se centra no solo los demás puntos que menciona el autor sino también el articulo se centra en este punto de vista, el cual menciona el autor “Vamsi K” que está en casi en total desacuerdo con la publicación anterior realizada de “KDnuggets” mencionando que Hadoop falla, por ello “Vamsi K” deja claro argumentando que Hadoop es una plataforma tecnológica robusta y madura que tiene mucho que ofrecer, de esta forma fundamenta a lo largo del articulo el porque de ese punto de vista y no solo con argumentos sino también por experiencia propia de su trabajo y conocimiento que posee al respecto.

***Punto2.*** Otro de los puntos de vista que le autor deja ver en el articulo es sobre su visión respecto a Hadoop y como este seguirá prosperando en los próximos años, al inicio lo respalda mencionando su gran experiencia en el área de Big data en el extranjero junto a los lideres bancarios y que debido a ello el sabe que Hadoop esta prosperando. Además, menciona que todos los desafíos como lo son la capacidad de trabajar a grandes velocidades y la capacidad de obtener información en tiempo real a grandes flujos de datos, Hadoop los convierte en oportunidades comerciales para adoptantes eficientes.

***Punto3.*** Un tercer punto de vista es que Hadoop es una plataforma tecnológica robusta y madura, donde el autor entrega su visión al respecto diciendo que su punto de vista es garantizado o lo que le da garantía a Hadoop ser lo que es gracias a el ecosistema de código abierto darwiniano ya que el área de Big Data tiene es respaldado por la comunidad de código abierto que tiene la mayor parte de la tecnología del ecosistema Hadoop, siendo esta comunidad inherentemente de naturaleza darwiniana por ello difícil que decaiga o que lo mencionado por “Vamsi K” este en lo correcto.

***Punto4.*** Como ultimo punto a explicar es mas bien un conjunto de visiones o argumentaciones que hace el autor respecto a las suposiciones que da “Vamsi K”, por ello explicare el punto de vista del autor que va en su mayoría en desacuerdo con las 4 suposiciones dadas por el otro autor “Vamsi K”. Lo primero a mencionar por el autor es que la adopción de Hadoop está lejos de estar estancada, de hecho se está disparando a áreas como la Banca, Telecomunicaciones, Manufactura y Seguros, por consiguiente el autor menciona numerosos logros o avances relacionados a esta expansión y crecimiento, además de realizar una comparativa en relación con la rapidez que lo hace Red Hat. Lo segundo que se menciona es que, si está de acuerdo en un punto mencionado por “Vamsi K” en relación al valor comercial de los proyectos creados con Hadoop y que estos no están claros, donde el autor explica que no es por que no este realmente claro y no es un punto en contra de Hadoop, sino que es algo lógico teniendo en cuenta que las organizaciones no solo utilizan Big Data para resolver desafíos sino también para experimentar nuevos modelos comerciales. Lo tercero a mencionar es que el autor esta en total desacuerdo con lo mencionado por “Vamsi K” y que esta fuera de la realidad el hecho que Big data solo aplica a soluciones válidas para volúmenes de Petabytes, por ello menciona que el real valor es poder construir repositorios de datos entre organizaciones que eran demasiados caros o con gran dificultad, además que muchos aprovechan la velocidad que entrega Big data mientras trabajan con TB y que se recomienda a los clientes comenzar por algo pequeño, además de tener almacenamiento económico, variedad de procesamiento de datos , entre otras ventajas mencionadas. Finalmente el ultimo punto al que el autor hace referencia es respecto a la dificultad de encontrar habilidades en Hadoop, donde el autor argumenta que es verdad y existe esa dificultad pero en áreas como lo son Data Scientist, Data Engineers y Hadoop Administrators, además de que no solo ocurre en Hadoop sino en todas las nuevas tecnologías y esto es superable en las empresas.